732A38 Computational statistics
Division of Statistics and Machine Learning
Department of Computer and Information Science

Computer lab 4

Instructions

- Create a report to the lab solutions in PDF.
- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report.**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1: Computations with Metropolis-Hastings

Consider the following probability density function:

$$f(x) \propto x^5 e^{-x}, \;\; x > 0$$

You can see that the distribution is known up to some constant of proportionality.

1. Use Metropolis-Hastings algorithm to generate samples from this distribution by using proposal distribution as log-normal LN(X_t, 1), take some starting point. Plot the chain you obtained as a time series plot. What can you guess about the convergence of the chain? If there is a burn-in period, what can be the size of this period?
2. Perform step 1 by using chi-square distribution $\chi^2(floor(X_t+1))$ as proposal distribution where floor(x) means integer part of *x*.
3. Compare the results of steps 1 and 2 and make conclusions.
4. Generate 10 MCMC sequences using the generator from the step 2 and with starting points 1,2,.., or 10. Use Gelman-Rubin method to analyze convergence of these sequences.
5. Estimate $\int_0^\infty \mathrm{x} \cdot \mathrm{f(x)dx}$ using the samples from steps 1 and 2.
6. The distribution generated is in fact a gamma distribution. Look in the literature and define the actual value of the integral. Compare it with the one you obtained.

Assignment 2: Gibbs sampling

A concentration of a certain chemical was measured in a water sample, and the result was stored in the data **chemical.RData** having the following variables:

- X: day of the measurement
- Y: measured concentration of the chemical.

The instrument used to measure the concentration had certain accuracy; this is why the measurements can be treated as noisy. Your purpose is to restore the expected concentration values.

1. Import the data to R and plot the dependence of Y on X. What kind of model is reasonable to use here?
2. A researcher has decided to use the following (random-walk) Bayesian model (*n*=number of observations, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ are unknown parameters):
$$Y_i \sim N(\mu_i, var = 0.2), i = 1, \dots, n$$
where the prior is
$$p(\mu_1) = 1$$
$$p(\mu_{i+1}|\mu_i) = N(\mu_i, 0.2), i = 1, \dots, n-1$$
Present the formulas showing the likelihood $p(\boldsymbol{Y}|\boldsymbol{\mu})$ and the prior $p(\boldsymbol{\mu})$ (hint: a chain rule can be used here $p(\boldsymbol{\mu}) = p(\mu_1)p(\mu_2|\mu_1)p(\mu_3|\mu_2) \dots p(\mu_n|\mu_{n-1})$)
3. Use the Bayes theorem to get the posterior up to a constant of proportionality, and then find out the distributions for $\mu_i|\boldsymbol{\mu}_{-i}, \boldsymbol{Y}$ where $\boldsymbol{\mu}_{-i}$ is a vector containing all μ values except of μ_i
   a. Hint A: consider separate formulas for $\mu_1|\mu_{-1}, \boldsymbol{Y}$, $\mu_{50}|\mu_{-50,}\boldsymbol{Y}$ and then a formula for all remaining $\mu_i|\boldsymbol{\mu}_{-i}, \boldsymbol{Y}$
   b. Hint B: $\exp\left(-\frac{1}{d}((x-a)^2 + (x-b)^2)\right) \propto \exp\left(-\frac{\left(x-\frac{a+b}{2}\right)^2}{d/2}\right)$
   c. Hint C: $\exp\left(-\frac{1}{d}((x-a)^2 + (x-b)^2 + (x-c)^2)\right) \propto \exp\left(-\frac{\left(x-\frac{a+b+c}{3}\right)^2}{d/3}\right)$
4. Use the distributions derived in step 3 to implement a Gibbs sampler that uses $\boldsymbol{\mu}^{\mathbf{0}} = (0, \dots, 0)$ as a starting point. Run the Gibbs sampler to obtain 1000 values of $\boldsymbol{\mu}$ and then compute the expected value of $\boldsymbol{\mu}$ by using Monte Carlo approach. Plot the expected value of $\boldsymbol{\mu}$ versus X and Y versus X in the same graph. Does it seem that you have managed to remove the noise? Does it seem that the expected value of $\boldsymbol{\mu}$ can catch the true underlying dependence between Y and X?
5. Make a trace plot for μ_{50} and comment on the burn-in period and convergence.

Submission procedure

Assume that X is the current lab number.

If you are neither speaker nor opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members does the following before the deadline:
  - submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
  - Goes to Study room *Speakers X→Documents* and opens file *Password X.txt*. Then the student should put your group report into ZIP file *Lab X.zip* and protect it with a password you found in Password X.txt
  - Uploads the file to *Collaborative workspace* folder

If you are opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
- After the deadline for the lab has passed, go to Collaborative workspace folder and download *Lab X.zip*. Open the PDF in this ZIP file by using the password available in *Course Documents→Password X.txt,* read it carefully and **prepare at least two questions/comments/improvement suggestions** in order to put them at the seminar (i.e. at least two questions per opponent)